**Homework 1**

**Azure Cloud Analytics**

1. **Please review Azure Concepts Section at follow link on Linux Academy:**

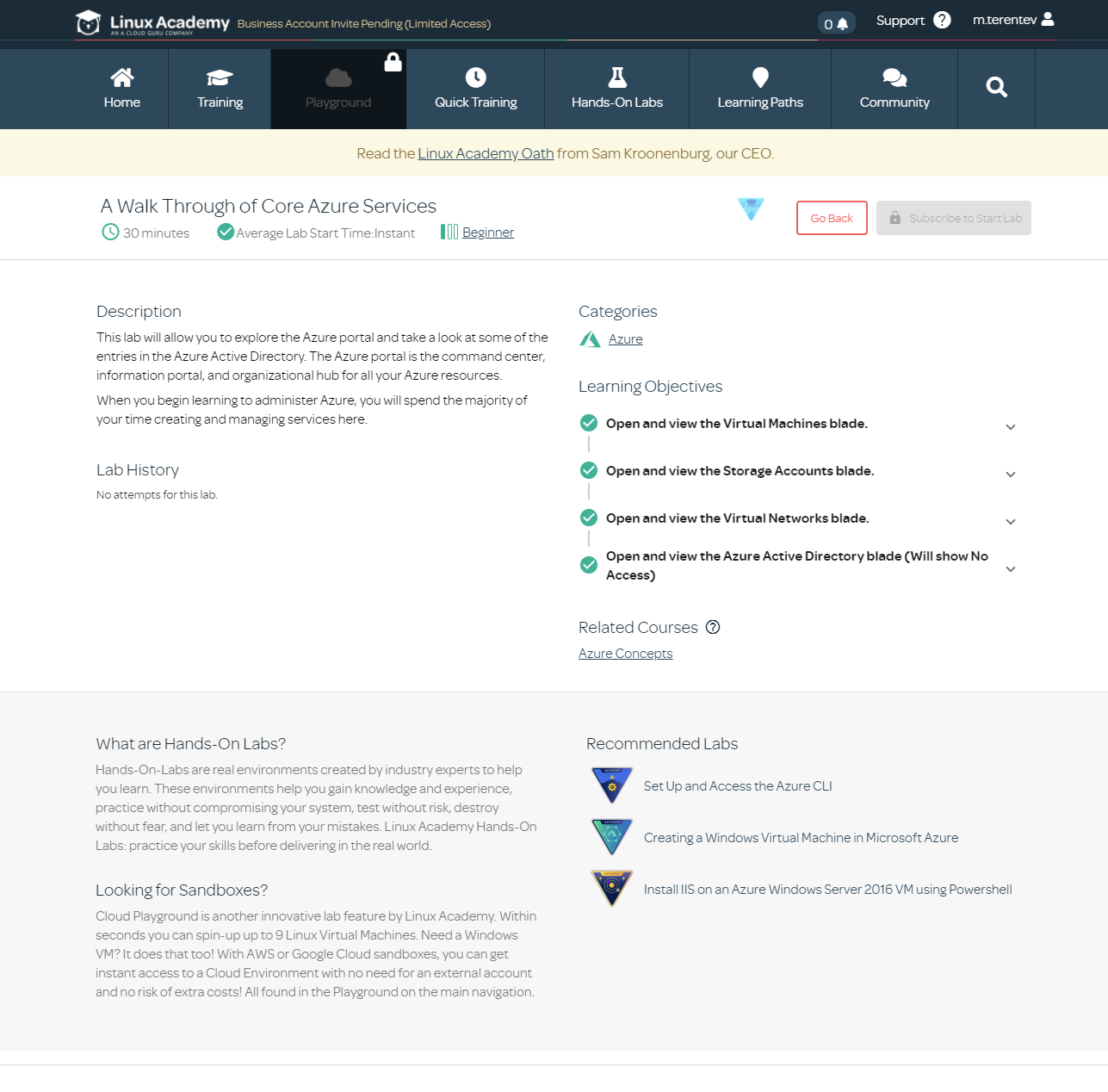
<https://linuxacademy.com/cp/modules/view/id/259>

Complete all items in this learning path on Linux Academy:

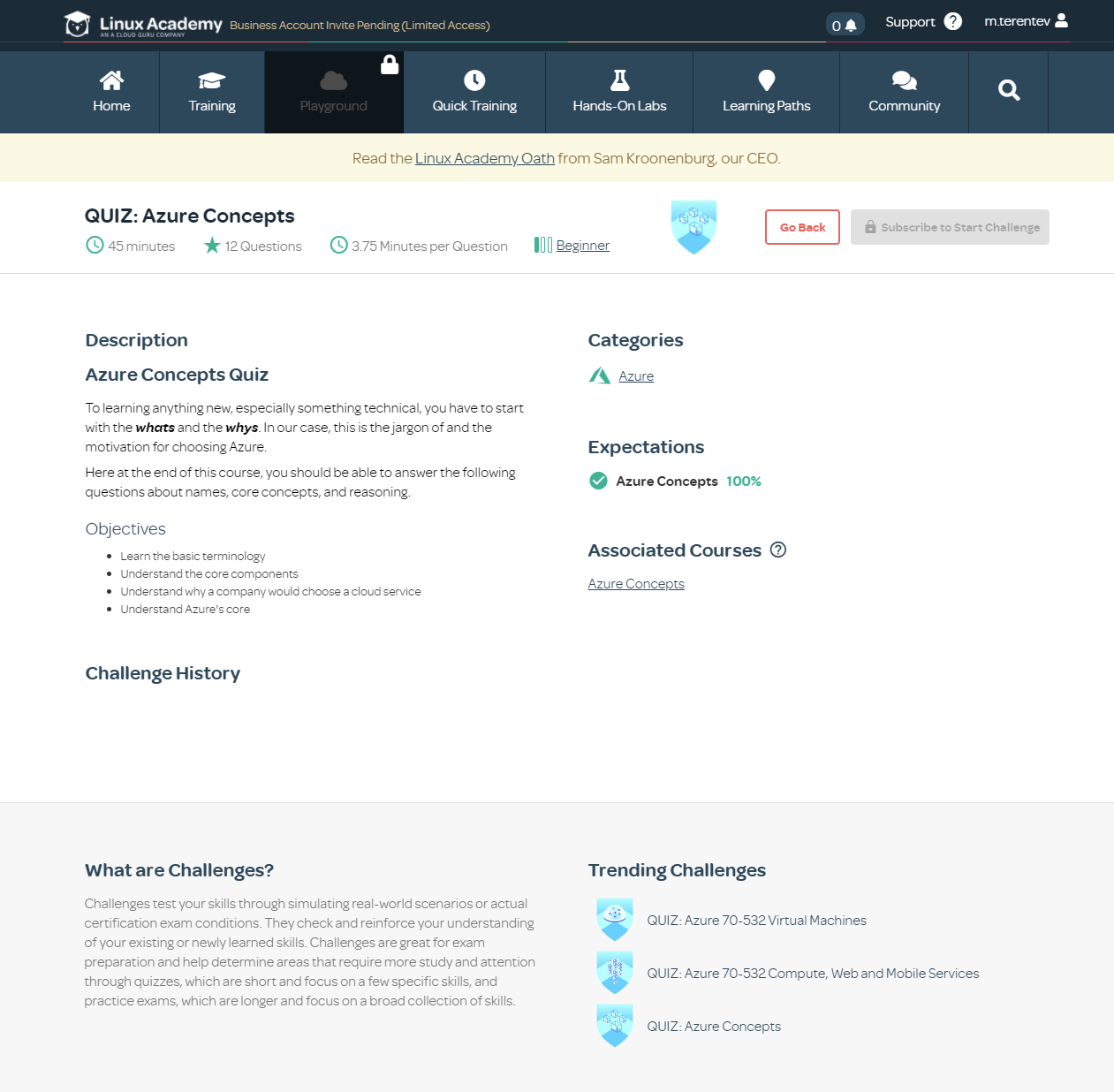
Complete the Hands On Lab associated with this section and complete the Quiz.

1. Explain in your own words your understanding on Azure components listed below:
2. High Availability, Fault tolerance
   1. High availability: technologies that minimize IT disruptions by providing business continuity of IT services through redundant, fault-tolerant, or failover-protected components inside the same data center.
   2. Fault tolerance: If any component fails on the primary replica, Azure detects the failure and fails over to the secondary replica.
3. Scalability and Elasticity
   1. Scaling: adaptability of the system to the changed amount of workload or traffic to the web application.
   2. Elasticity: the ability to increase or decrease in capability as workload demand changes.
4. Active Directory
   1. Azure Active Directory: Microsoft's cloud-based identity and access management service, which helps users sign in and access resources in.
5. Virtual Machine
6. Virtual machine: sandboxed from the rest of the system, meaning that the software inside a virtual machine cannot escape or tamper with the computer itself. Azure provide two types of VM: Windows and Linux-based.
7. Virtual Networks
   1. Virtual Networks: a representation of isolated networks in the cloud. It is a logical isolation of the Azure cloud dedicated to subscription.
8. Azure Storage
   1. Azure Storage: Secure cloud storage and provides with control over who has access to data.
9. Complete Hands On Lab using privileges provided on Azure Concepts section and explain what you understood from this lab.

I am unable to take this Lab

1. Complete the Quiz and Put screenshot of your score with your homework submission.

I am unable to take this Quiz, the ‘Unlock challenge’. I have completed the course.



1. Read following Microsoft Azure Architecture articles and summarize your understanding on each of the items listed below:
   1. Run Windows Virtual Machine in Azure: <https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/n-tier/windows-vm>

This article shows best practices for running a Windows VM on Azure. It requires some additional components besides the VM itself, including networking and storage resources such as disks. It also provides the considerations for cost, security, and DevOps. Below is the summary for the best practices.

Resource group: Group resources based on their lifetime and who will manage them.

Virtual Machine: Provision a VM from a list of published images, or from a custom managed image or virtual hard disk (VHD) file uploaded to Azure Blob storage. Generally, choose an Azure region that is closest to your internal users or customers.

Disks: When you provision a disk, consider all three factors (capacity, IOPS, and throughput).

Network: The networking components include the following resources: virtual network, network interface (NIC), Public IP address, static IP address, a fully qualified domain name (FQDN), and Network security group (NSG).

Operations: Diagnostics, availability, backups, stopping a VM, and deleting a VM.

Cost considerations: Consider different options for Pay as you go, Azure reservations, Azure Spot VMs. Use the Azure Pricing Calculator to estimates costs.

Security considerations: Use Azure Security Center to get a central view of the security state of your Azure resources including patch management, antimalware, access control, audit logs, and data encryption.

DevOps considerations: Use infrastructure as Code (IaC) for provisioning the Azure resources to perform continuous integration and continuous delivery (CI/CD).

This article shows best practices for running a Windows VM on Azure. It requires additional com

* 1. Information discovery with deep learning and NLP: <https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/information-discovery-with-deep-learning-and-nlp>

This article provides the information of architecture and components for deep learning and natural language processing (NLP). By combining deep learning and NLP with data on site-specific search terms, this solution helps greatly improve tagging accuracy on social sites.

* 1. Mass ingestion and analysis of news feeds on Azure: <https://docs.microsoft.com/en-us/azure/architecture/example-scenario/ai/newsfeed-ingestion>

This article describes mass ingestion and near real-time analysis of documents using public RSS news feeds. It uses Azure Cognitive Services to offer useful insights including text translation, facial recognition, and sentiment detection. You can extend the news feeds to other languages besides English.

This article provides the use cases, architecture, consideration, pricing, and codes from GitHub to deploy the scenario.

* 1. Azure data platform end-to-end: <https://docs.microsoft.com/en-us/azure/architecture/example-scenario/dataplate2e/data-platform-end-to-end>

This article demonstrates how to use the Azure Data Services to build a modern data platform capable of handling the most common data challenges in an organization. The solution combines a range of Azure services that will ingest, process, store, serve, and visualize data from different sources, both structured and unstructured.

It provides information for relational databases, semi-structured and non-structured data sources, and steaming. The article provides the considerations, pricing, and alternatives option.

1. Read through all Design patterns and explain each Design Pattern in few lines based on your understanding

Ambassador is Create helper services that send network requests on behalf of a consumer service or application.

Anti-Corruption Layer is Implement a façade or adapter layer between a modern application and a legacy system.

Asynchronous Request-Reply is Decouple backend processing from a frontend host, where backend processing needs to be asynchronous, but the frontend still needs a clear response.

Backends for Frontends is Create separate backend services to be consumed by specific frontend applications or interfaces.

Bulkhead is Isolate elements of an application into pools so that if one fails, the others will continue to function.

Cache-Aside is Load data on demand into a cache from a data store

Choreography is Let each service decide when and how a business operation is processed, instead of depending on a central orchestrator.

Circuit Breaker is Handle faults that might take a variable amount of time to fix when connecting to a remote service or resource.

Claim Check is Split a large message into a claim check and a payload to avoid overwhelming a message bus.

Compensating Transaction is Undo the work performed by a series of steps, which together define an eventually consistent operation.

Competing Consumers is Enable multiple concurrent consumers to process messages received on the same messaging channel.

Compute Resource Consolidation is Consolidate multiple tasks or operations into a single computational unit

CQRS is Segregate operations that read data from operations that update data by using separate interfaces.

Deployment Stamps is Deploy multiple independent copies of application components, including data stores.

Event Sourcing is Use an append-only store to record the full series of events that describe actions taken on data in a domain.

External Configuration Store is Move configuration information out of the application deployment package to a centralized location.

Federated Identity is Delegate authentication to an external identity provider.

Gatekeeper is Protect applications and services by using a dedicated host instance that acts as a broker between clients and the application or service, validates and sanitizes requests, and passes requests and data between them.

Gateway Aggregation is Use a gateway to aggregate multiple individual requests into a single request.

Gateway Offloading is Offload shared or specialized service functionality to a gateway proxy.

Gateway Routing is Route requests to multiple services using a single endpoint.

Geodes is Deploy backend services into a set of geographical nodes, each of which can service any client request in any region.

Health Endpoint Monitoring is Implement functional checks in an application that external tools can access through exposed endpoints at regular intervals.

Index Table is Create indexes over the fields in data stores that are frequently referenced by queries.

Leader Election is Coordinate the actions performed by a collection of collaborating task instances in a distributed application by electing one instance as the leader that assumes responsibility for managing the other instances.

Materialized View is Generate prepopulated views over the data in one or more data stores when the data isn't ideally formatted for required query operations.

Pipes and Filters is Break down a task that performs complex processing into a series of separate elements that can be reused.

Priority Queue is Prioritize requests sent to services so that requests with a higher priority are received and processed more quickly than those with a lower priority.

Publisher/Subscriber is Enable an application to announce events to multiple interested consumers asynchronously, without coupling the senders to the receivers.

Queue-Based Load Leveling is Use a queue that acts as a buffer between a task and a service that it invokes in order to smooth intermittent heavy loads.

Retry is Enable an application to handle anticipated, temporary failures when it tries to connect to a service or network resource by transparently retrying an operation that's previously failed.

Scheduler Agent Supervisor is Coordinate a set of actions across a distributed set of services and other remote resources.

Sequential Convoy is Process a set of related messages in a defined order, without blocking processing of other groups of messages.

Sharding is Divide a data store into a set of horizontal partitions or shards.

Sidecar is Deploy components of an application into a separate process or container to provide isolation and encapsulation.

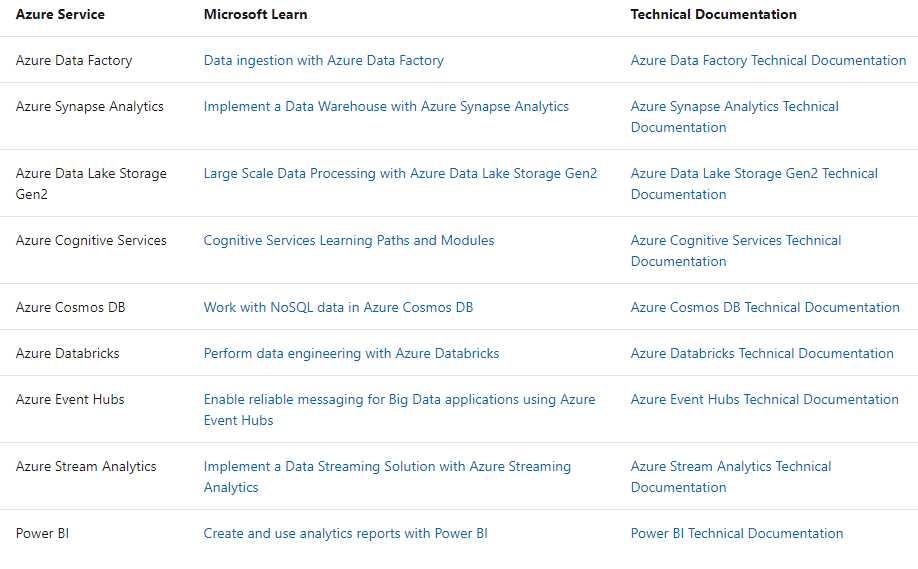
Static Content Hosting is Deploy static content to a cloud-based storage service that can deliver them directly to the client.

Strangler is Incrementally migrate a legacy system by gradually replacing specific pieces of functionality with new applications and services.

Throttling is Control the consumption of resources used by an instance of an application, an individual tenant, or an entire service.

Valet Key is Use a token or key that provides clients with restricted direct access to a specific resource or service.

1. Read through following white paper on End to End Data Platforms on Azure: <https://docs.microsoft.com/en-us/azure/architecture/example-scenario/dataplate2e/data-platform-end-to-end>
   1. Explain each of the following Data Platforms. Focus on Function, use case and applications for each.



Azure Data Factory: Azure Data Factory is Azure's cloud extract-transform-load (ETL) service for scale-out serverless data integration and data transformation. Data Factory contains a series of interconnected systems that provide a complete end-to-end platform for data engineers.

Azure Synapse Analytics: Azure Synapse is an analytics service that brings together enterprise data warehousing and Big Data analytics. It gives you the freedom to query data on your terms, using either serverless on-demand or provisioned resources—at scale.

Azure Data Lake Storage Gen2: Azure Data Lake Storage Gen2 is a set of capabilities dedicated to big data analytics, built on Azure Blob storage. Data Lake Storage Gen2 is the result of converging the capabilities of our two existing storage services, Azure Blob storage and Azure Data Lake Storage Gen1.

Azure Cognitive Services: Azure Cognitive Services are cloud services with REST APIs, and client library SDKs available to help developers build cognitive intelligent applications without having direct artificial intelligence (AI) or data science skills or knowledge.

Azure Cosmos DB: Azure Cosmos DB is Microsoft’s globally distributed, multi-model database service for operational and analytics workloads. It offers multi-mastering feature by automatically scaling throughput, compute, and storage.

Azure Databricks: Azure Databricks is an Apache Spark-based analytics platform optimized for the Microsoft Azure cloud services platform. Designed with the founders of Apache Spark, Databricks is integrated with Azure to provide one-click setup, streamlined workflows, and an interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.

Azure Event Hubs: Azure Event Hubs is a big data streaming platform and event ingestion service. It can receive and process millions of events per second. Data sent to an event hub can be transformed and stored by using any real-time analytics provider or batching/storage adapters.

Azure Stream Analytics: Azure Stream Analytics is a real-time analytics and complex event-processing engine that is designed to analyze and process high volumes of fast streaming data from multiple sources simultaneously. Patterns and relationships can be identified in information extracted from a number of input sources including devices, sensors, clickstreams, social media feeds, and applications.

Power BI: Power BI is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights. Your data may be an Excel spreadsheet, or a collection of cloud-based and on-premises hybrid data warehouses. Power BI lets you easily connect to your data sources, visualize and discover what's important, and share that with anyone or everyone you want.

* 1. Explain difference between Relational, Unstructured, Streaming and semi structured data.

Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns all data which can be stored in database SQL in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data are most processed in the development and simplest way to manage information. Example: Relational data.

Semi-structured data is information that does not reside in a relational database but that have some organizational properties that make it easier to analyze. With some process, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. Example: XML data.

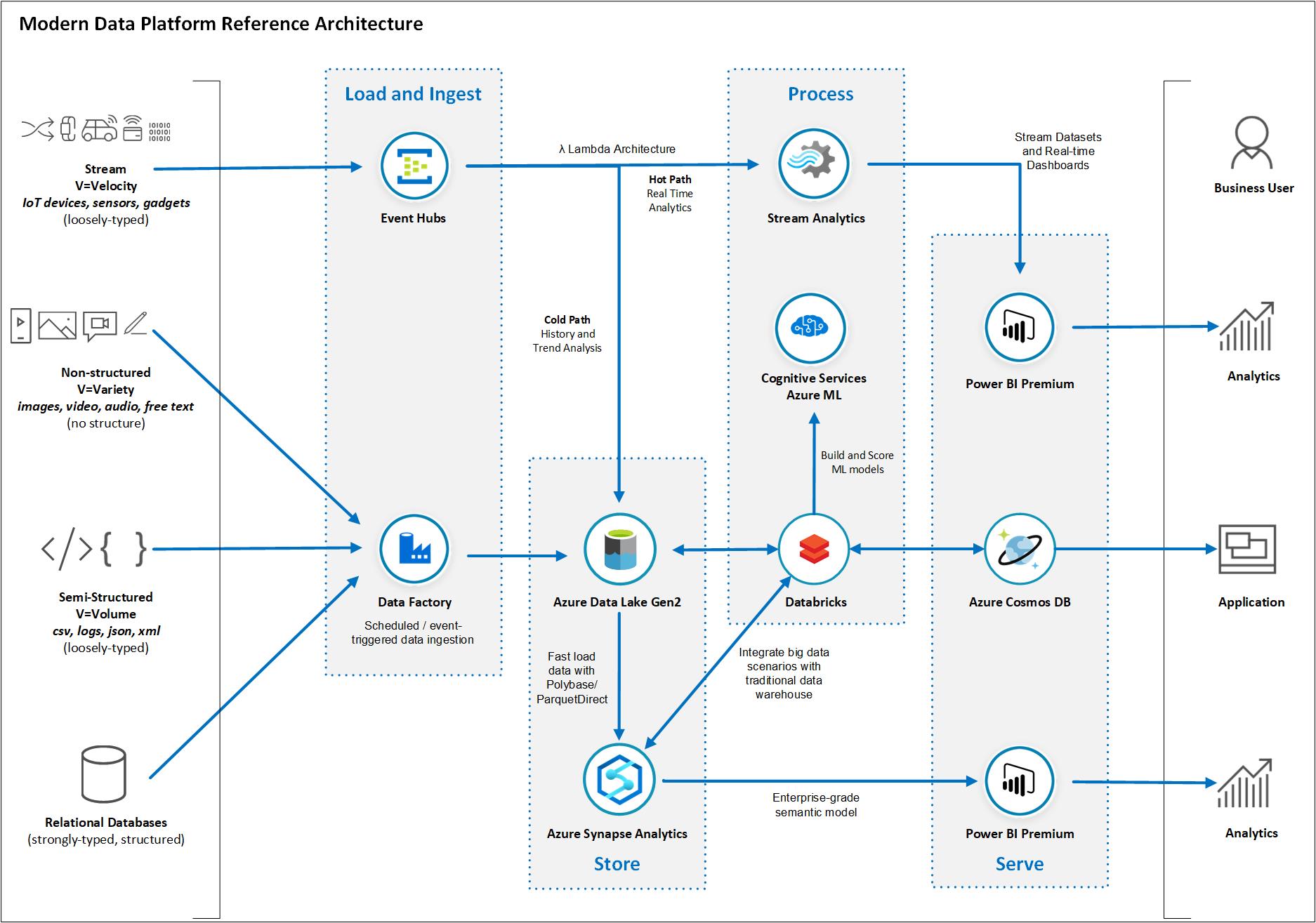
Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. Example: Word, PDF, Text, Media logs.

* 1. Explain Analytical Data Store in Azure <https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/analytical-data-stores>

This article talks about the options when choosing an analytic data store. The options are:

* Azure Analysis Services
* Azure Cosmos DB
* Azure SQL Database
* Azure Synapse Analytics
* HBase/Phoenix on HDInsight
* Hive LLAP on HDInsight
* SQL Server in Azure VM

Also, it provides the key selection criteria and capability matrix.

* 1. Please note following view graph and explain following tools in few lines:

Explain key differences, function and key features of each of the following tools in Azure data platform.

* + 1. Databricks

Microsoft Azure Designed with the founders of Apache Spark, Databricks is integrated with Azure to provide one-click setup, streamlined workflows, and an interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.

* + 1. Azure Synapse

Azure Synapse is a limitless analytics service that brings together enterprise data warehousing and Big Data analytics.

* + 1. Azure Data Lake Gen2

Azure Data Lake includes all the capabilities required to make it easy for developers, data scientists, and analysts to store data of any size, shape, and speed, and do all types of processing and analytics across platforms and languages.

* + 1. Data Factory

Data Factory designed service to allow developers to integrate disparate data sources.  It is a platform somewhat like SSIS in the cloud to manage the data you have both on-prem and in the cloud.

* + 1. Azure Cosmos DB

Azure Cosmos DB is a fully managed NoSQL database service for modern app development with guaranteed single-digit millisecond response times and 99.999-percent availability.